ABC 2021, 8(2):152-165
Animal Behavior and Cognition
DOI: https://doi.org/10.26451/abc.08.02.05.2021

How Subtle Protocol Choices Can Affect Biological Conclusions: Great Tits' Response to Allopatric Mobbing Calls

Ambre Salis*, Jean-Paul Lena, & Thierry Lengagne

[1] Univ Lyon, Université Claude Bernard Lyon 1, CNRS, ENTPE, UMR 5023 LEHNA, F-69622, Villeurbanne, France

*Corresponding author (Email: ambre.salis@univ-lyon1.fr)

Citation – Salis, A., Lena, J-P., & Lengagne, T. (2021). How subtle protocol choices can affect biological conclusions: Great tits' response to allopatric mobbing calls. *Animal Behavior and Cognition*, *8*(*2*), 152-165. https://doi.org/10.26451/abc.08.02.05.2021

Abstract – In the last ten years, numerous replicated studies showed divergent results from the original papers, leading to the recognition that science may be facing a replication crisis. Apart from fraud or natural population variability, different results may emerge from flexibility in the protocol and/or restricted sample sizes. Replication studies are therefore fundamental to assess how robust a finding can be. However, while the original authors may be prone to p-hacking (to collect data, select data or use statistical analyses until nonsignificant results become significant), the replication-authors are, on the contrary, probably unwittingly prone to show opposite results (i.e., null-hacking). In this study, we face the unique opportunity to compare replicated studies with no null-hacking bias. Two teams independently investigated the response of great tits (*Parus major*) to mobbing calls of an allopatric species, in their natural and reversed order, on the same population. We first discuss how subtle protocol choices, especially regarding the soundtrack preparation and playback methodology, can explain variation in the results. In addition, we show that, although the effect sizes of the differences of interest are similar, biological conclusions solely based on the p-value would be different. Thus, we note the pitfall of reliance on p values, especially with small samples.

Keywords – Animal communication, Heterospecific communication, Mobbing, Replication crisis, Syntax

During the last decade, failure to reproduce published results in various fields of research (e.g., in psychology: Bohannon, 2015, or in epidemiology: Lash et al., 2018) alerted the scientific community about a potential low reliability of published results. This replication crisis, complex and heavily debated, can be explained in different ways (Fanelli, 2018; Maxwell et al., 2015). Indeed, while models have shown that the global "publish or perish" problem probably increases misconduct (Grimes et al., 2018; Higginson & Munafò, 2016), direct frauds (i.e., fabrication or falsification of data) remain scarce according to empirical evidence (Fanelli, 2018). More probable is the effect of inconspicuous and ordinary factors (Ioannidis, 2005) leading to the publication of unreliable results and/or interpretations, with one major example being conscious but also unconscious p-hacking (i.e., to collect data, select data or use statistical analyses until nonsignificant results become significant, Head et al., 2015).

In the field of animal behavior and especially in animal communication, three specific factors can increase the publication of contrasting results. Firstly, behavior is, by nature, an external proxy of internal

states of animals, needing researchers to interpret each variable under study. Such behaviors are prone to flexibility between researchers regarding the protocol used, definition of the behaviors, and their relative relevance to the question asked. Indeed, each scientist is the sum of their past experience, knowledge, and personal background, which can affect their conclusions at the creation of the protocol (Tang-Martínez, 2020), when analyzing results (Silberzahn et al., 2018) or when interpreting those results (Tang-Martínez, 2020). Greater flexibility, in interaction with natural population variability (Danchin et al., 2008), increases the opportunity to transform negative results into positive ones (Ioannidis, 2005), and, in a general manner, creates disparities between papers investigating the same question. Secondly, the difficulty to obtain large numbers of wild subjects as much as the ethics considerations often lead behavior studies to obtain restricted sample sizes (Schwagmeyer & Mock, 1997). Small sample sizes are less likely to detect small differences between treatments (type II error, Button et al., 2013) but are also prone to stochastic variation so that the probability of a positive result is inflated despite no biological difference (type I error; Button et al., 2013). Thirdly, fields of research such as language evolution in animal communication are quite new, with several teams working simultaneously on similar questions. This increases the risk of more spectacular positive results being published in priority, as each team aims at showing their most influential work (Ioannidis, 2005).

Replicating behavioral studies should consequently be of great interest, with one caveat: while the original author may have been prone to *p*-hacking, the replicating author may in opposition (probably unwittingly) possess a "null hacking bias" (i.e., the motivated pursuit of null results by replicating investigators, Bryan et al., 2019). As a result, replication studies are often as questionable as the study they wish to replicate (Schmidt & Oh, 2016), and cannot alone be sufficient to conclude on the biological question at stake. To circumvent this problem, one would need two researchers to blindly replicate a study (i.e., having the same question, on the same population, without being influenced by each other). Such a situation occurred in our laboratory: two independent researchers, one leaving and one arriving in the laboratory, had by chance the same idea, and communicated too late about it. This led to two datasets answering the same question, obtained with quite similar yet not exactly equal protocols. Because the disparities between protocols are relatively low, this presented one great opportunity to investigate how flexibility in the protocol and limited sample sizes can affect, or not, the resulting biological conclusions.

In this article, we will therefore compare our work (presented for the first time here) to the work of Dutour et al. (2020), both investigating the impact of reversed syntax on response of great tits to heterospecific calls (see *§Biological question* for details). We will not discuss the importance of the resulting biological conclusions regarding compositional syntax in birds, since Dutour and her colleagues (2020) already did so. We will focus on whether slight differences in protocol choices resulted in contrasting results, discuss which parameters may be of importance in such potential disparities and conclude on how this affects our field of research.

Method

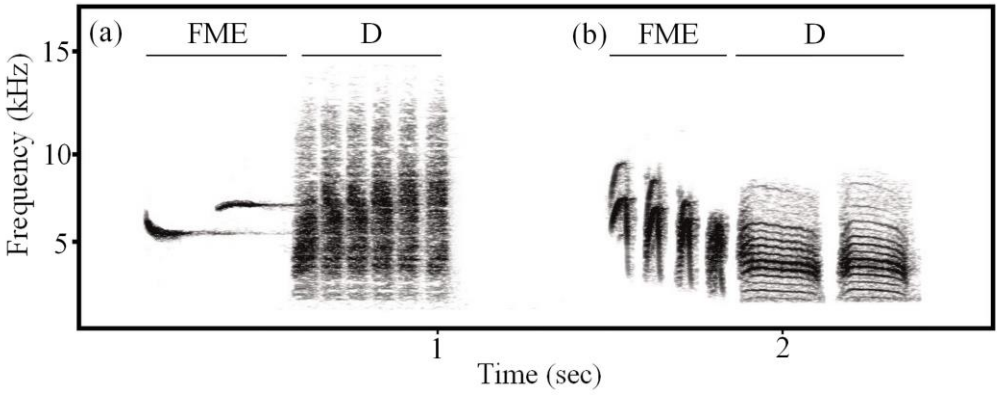
Biological Question

The experiments independently carried out by both teams concern the currently hotly debated question of compositional syntax in birds. Compositional syntax is defined as when the meaning of a sequence is related to its different parts and in the way they are combined (Suzuki et al., 2019, 2020). Recent studies have proposed that some species, when mobbing a predator (i.e., actively harass it instead of flying away, Carlson et al., 2018), use a combinatorial call in a fixed order: the first part (hereafter called FME: Frequency Modulated Elements, Dutour et al., 2017) elicits vigilance, while the second part (called the D notes, following Hailman et al., 1985) elicits approach from the receiver (Dutour, Lengagne et al., 2019; Suzuki et al., 2016, Figure 1). Combined, the resulting sequence evokes behaviors such as scanning, approaching and calling in receivers, typical behaviors linked to mobbing (Carlson et al., 2017; Salis et al., 2021; Suzuki et al., 2016). Furthermore, the reversed order (i.e., D notes then FME) results in lower responses from the birds (Dutour, Lengagne et al., 2019; Suzuki et al., 2016). Debates on whether

such coding strategies can be designated as compositional syntax in the human linguistics sense have been profuse (Bolhuis et al., 2018a, b; Griesser et al., 2018).

Figure 1

Spectrograms of a Typical Mobbing Call of (a) Great Tits (Parus major), and (b) Black-capped Chickadees (Poecile atricapillus)



The reasonable response to such critics is that this young subject deserves more studies on the same species to conclude on such potentially high cognitive abilities in birds. One way to dig into that question can be to test whether a species known to use an FME-D combinatoriality also responds to mobbing calls of an allopatric species exhibiting a similar ordering sequence in mobbing call but made up of acoustically different notes. Two adequate species for such an experiment are the great tit (*Parus major*), living in Europe and for which the use of compositional syntax has already been investigated (Dutour, Lengagne et al., 2019), and the North American black-capped chickadee *(Parus atricapillus*), for which the mobbing calls are also made up of a FME-D notes combination with fixed syntactic rules (although the idea that FME notes are related to vigilance behavior and D notes to approach has not been tested yet in the black-capped chickadee, Baker & Becker, 2002; Otter, 2007, Figure 1). Great tits respond to the mobbing calls of the black-capped chickadee in the same way they do for conspecific calls: they mob to the complete sequence, are vigilant when hearing FME notes, and approach to D notes (Randler, 2012; Salis et al., 2021). One could therefore expect great tits to respond with mobbing behavior when presented with an unknown call sequence when it has the same composition, while failing to do so when the ordering of the call sequence is reversed, as they did for conspecific calls (Dutour, Lengagne et al., 2019). Such questions were addressed by two studies conducted within a few months of each other (Dutour et al., 2020 and the present study).

Field Experimental Protocol

Our experiment is aimed at answering two specific questions: (i) do great tits respond to allopatric mobbing sequences never heard before in the same way as they do for conspecific calls (question 1, hereafter designated as the "species comparison"), and (ii) would they do so for allopatric calls for which order is reversed (i.e., D-FME, question 2, hereafter designated as "order comparison").

To do so, a in a field study, we presented great tits with mobbing recordings of great tits, natural calls of black-capped chickadees, reversed calls of black-capped chickadees, or background noise (control). We measured their vigilance with the number of scans they produced, and whether they approached the loudspeaker. We hypothesized that they should scan and approach as if faced with conspecific calls, although the overall level of response may be reduced. Indeed, Randler (2012) found reduced response toward black-capped chickadees' mobbing calls, but Dutour et al. (2017) found similar level of response between conspecific and heterospecific calls. Secondly, if order is important in the

decoding process, they should not respond anymore when the allopatric mobbing sequence is reversed (less scanning and less approaching).

We here describe our protocol and for each point, describe the similarities or difference with Dutour et al. (2020). Our protocols are similar on most points but, while Dutour et al. (2020) created two separate experiments with somewhat different protocols (see Table 1), all our different treatments were tested in the same global experiment. We will consequently separate the two questions (species comparisons and order comparison) only in the statistical analysis and results section. For clarity's sake, Table 1 summarizes the common ground and differences with Dutour et al. (2020).

Table 1

Protocol Comparison Between the Experiment of M. Dutour and Colleagues and A. Salis and Colleagues

Protocol Choices	Salis et al. (current paper); Question 1 & 2	Dutour et al. (2020); Question 1	Dutour et al. (2020); Question 2
Receiver species	Great tit	Great tit	Great tit
Emitter species	Black-capped chickadee	Black-capped chickadee	Black-capped chickadee
Date	April-May 2019	February-March 2018	May 2018
Location of tests	North of Lyon, France	North of Lyon, France	North of Lyon, France
Bird tested	Free ranging birds	Free ranging birds	**Birds at nest boxes**
Number of treatments	4	2	3
N per treatment	20	20	20 (repeated measures)
Experimental design	CRD	CRD	**Crossover design**
Distance sampling (m)	100	100	50 (nest boxes)
Control(s)	Background noise	Ø	Background noise
Soundtracks' origin	Xeno-canto + Macaulay Library	Xeno-Canto + own recordings	Xeno-Canto + Macaulay Library
Control for Number of Notes or Call length?	**Call length**	Number of D notes per call	Number of D notes per call
Total Duty cycle (/min)	20 sec for all playbacks	GT: 25.6 sec, BC: 28.9 sec	GT: 25.6 sec, BC: 28.9 sec
Number of FME notes/call	GT: 2, BC: 4	GT: 2, BC: 4	GT: 2, BC: 4
Number of D notes/call	GT: 6-8, BC: 2-3	8 for all playbacks	8 for all playbacks
Call length	0.80 sec for all playbacks	GT: 1 sec, BC: 2 sec	GT: 1 sec, BC: 2 sec
Call repetition (/min)	30	GT: 26 BC: 14	GT: 26 BC: 14
Distance with the loudspeaker	16 m (Approach = 8 m)	**30 m (Approach = 15 m)**	20 m (Approach = 10 m)
Double blind observation	**Yes (headphones)**	Partial (unaware but can hear the playback)	Partial (unaware but can hear the playback)
Variables of interest	Scan + Approach	Scan + Approach	Scan + Approach
Statistical analysis	GLM; Poisson & Binomial	GLM; Quasi Poisson & Binomial	GLM; Quasi Poisson & Binomial

Note. Experiments consisted in recording behavior of birds when hearing specific soundtracks. We listed the factors that could potentially influence different results in the two studies. Bold text emphasizes the differences between protocols that are targeted in the Discussion section. CRD = Completely randomized design, GLM = generalized linear model, GT= great tit, BC = black-capped chickadee.

Preparation of Soundtracks

In our experiment and in Dutour et al. (2020), four different types of soundtracks were built: first, soundtracks with the complete (FME-D) mobbing call sequence of the great tit (GT) or black-capped chickadee (BC), to check whether great tits responded in a similar way to allopatric calls and conspecific ones. Secondly, we built artificially reversed black-capped chickadee sequences (D-FME) to test the importance of order. At last, we both used a control, background noise (BN).

Our soundtracks of great tits and black-capped chickadees were built using recordings obtained from the Xeno-canto online database (www.xeno-canto.org) and the Macaulay Library (www.macaulaylibrary.org). Dutour et al. (2020) used the same websites (recordings previously used in

other studies) in addition to three recordings of their own of great tits. For both species, we conserved only good quality recording files (A or B grades) under the denomination "Alarm call"/"Call." Then, to ensure that the selected recordings truly represented mobbing call sequences, several features were controlled: first, in both studied species, a mobbing call corresponds to an association of FME and D notes (Figure 1). Hence, selected recordings were all made of the same FME and D notes reported in Baker and Becker (2002) and Templeton et al. (2005) for the black-capped chickadee and in Randler (2012) and Kalb et al. (2019b) for the great tit. In addition, for both species, the D notes' length is known to vary with the context (Kalb et al., 2019b; Templeton et al., 2005); we therefore checked that our soundtracks had the same length as the D notes used in mobbing calls ($X = 0.05 \pm 0.01$ sec for the great tit, $X = 0.18 \pm 0.02$ sec for the black-capped chickadee, mean ± standard deviation). Finally, for the great tit, we verified that no FME used in food or flight related contexts were the most predominant in any of our recordings (i.e., G, H, I and M notes associated with food for the great tit, Kalb et al., 2019a).

From these recording files, we built 40 soundtracks of 1 min mobbing sequences of great tits and black-capped chickadees (20 for each species, each provided from a different emitter) using Avisoft-SASLab software (files were converted into a wav format). To allow comparison between black-capped chickadees' and great tits' responses, we constructed every mobbing sequence with a similar duty cycle (~20 s of sound/min, Landsborough et al., 2019) and mobbing calls repetition (30 calls/min, natural range of repetition rate, Suzuki et al., 2016). Consequently, each mobbing call emitted by both species had a total similar FME duration (0.31 ± 0.06 s/call, mean ± SD) and D duration (0.50 ± 0.07 s/call), but not the same number of notes, since black-capped chickadees' D notes are longer (Figure 1). Dutour et al. (2020) created the same number of playbacks but chose to control for the number of D notes per call (8 D notes/call) and global duty cycle (25.6 s/min for the great tit and 28.9 s/min for the black-capped chickadee) instead of the number of calls/min. Therefore, the number of calls per playback was lower in the black-capped chickadee playbacks (14 calls/min) compared to the playbacks of great tits (26 calls/min). In both our experiment and Dutour et al. (2020), reversed playbacks were constructed by putting the FME notes after the D notes. We made sure that the space between the FME and D notes was the same before and after the manipulation. The reversed playbacks therefore possessed the exact same duty cycle and rhythm as the natural order playbacks. We also constructed 20 background noise soundtracks extracted from the original recordings (control soundtrack hereafter referred as BN). Each of these 80 soundtracks were cleared of any other bird calls, background noise was reduced, and amplitude homogenized.

Field Tests

In our experiment, data were collected in the east of France during the breeding season (March/April 2019), in a radius of 25 km around Lyon. Data of Dutour et al. (2020) were collected in the same territory, also during the breeding season, but in 2018 (February/March for question 1, May for question 2). In this region, great tits begin to defend territories in February, build their nests in March, and lay eggs at the beginning of April (personal observation). They usually lay 5 to 12 eggs, and then incubate for ~13days (Géroudet & Robert, 1954). Finally, nestlings are fed for ~19 days (Dutour, Léna et al., 2019), until their departure around mid-May.

For each type of soundtrack, 20 fully independent tests were performed (each bird tested was tested only once, and each bird received a different playback). In both Dutour et al. (2020) and our experiment, each test was performed by two field assistants. One of them was assigned to the soundtrack operation, while the other was kept unaware of the selected soundtrack (using headphones with music) and assigned to the behavioral recording of the focal bird. For each test, after detecting an individual using binoculars, the focal bird was observed for at least 1 min, and the pre-test behavior (singing or foraging) was noted. If the animal was displaying an alarm behavior, no test was performed. A loudspeaker was placed 16 m from the bird (16.79 ± 6.27 m), and at less than 3 m to a potential roost (bushes/trees) to allow the approach of the focal bird. The two field assistants were then placed in retreat (minimum of 15 m to both the bird and the loudspeaker) before launching the soundtrack with a remote

control. All soundtracks were broadcast using a Shopinnov 20 W loudspeaker with an intensity of 79.8 ± 1.9 dB(C) (measured at 1 m from the loudspeaker using Lutron SL-4001, C weighting, slow settings, re: 20 µPa). The field procedure for question 1 of Dutour et al. (2020) and in our experiment were both based on a complete randomized design and very similar, excepted for three details: 1) the main observer in Dutour et al. (2020) was aware of the playback launched, as she did not wear any sound protection, 2) the loudspeaker was placed at ~ 30 m from the bird in Dutour et al. (2020), and 3) the amplitude of the sound was of 83 dB in Dutour et al. (2020). In Experiment 2 of Dutour et al. (2020), the tests were carried out at the nest, the loudspeaker was placed at 20 m, and birds were tested several times using a crossover design.

Tests were carried out between 06:00 and 13:00 h during calm and dry weather days. Each of the four soundtracks were tested each day in a different order to avoid any temporal effect. To avoid pseudoreplication, each selected focal bird was separated from each other by at least 100 m (Dutour, Lengagne et al., 2019). Although birds were not individually ringed, great tits are known to be strongly territorial during the breeding period (Krebs, 1971; Wilkin et al., 2006) so that spacing between neighboring individuals is often used to ensure sampling of different specimens in field tests. As in several other studies (e.g., Dutour, Lengagne et al., 2019), we used a distance that roughly corresponded to the highest average distance expected according to territorial sizes reported in this species (c.a. 1.5 ha, Wilkin et al., 2006). Moreover, in the present study, two or three singing birds were often concurrently detected within 100 m, suggesting territorial size to be substantially inferior to 1.5 ha in the study area. Yet, we recognize that future experiments should separate tests of non-ringed birds by at least 250-300 m to ensure data independency.

Behavioral Observations

In our experiment and Dutour et al. (2020), during 1 min of playback, two types of behavioral states were assessed, respectively: 1) Vigilance effort as indicated by the number of horizontal scans displayed (the number of movements that birds made with their heads from left to right or right to left, approximately a 180° turn (Dutour, Lengagne et al., 2019; Suzuki et al., 2016); 2) Approach inferred using a dichotomic variable (approaching at least halfway from their starting point) measured with a Leica DISTO D210 telemeter. In the field, we reported, for each test, the distance of the bird from the loudspeaker at the beginning of the test, and the closest distance of the bird during the test. We then divided the closest distance by the distance at the beginning and the bird was considered as approaching when the ratio was < 0.5. This way of defining approach allowed us to take into account the initial distance of the bird (even if we tried to be at 16 m from the bird, we sometimes were at 14 m or 18 m). All observations were done using binoculars and recorded on a voice recorder (Sony ICD-PX370) by the same pair of observers in Dutour et al. (2020) while two trained pairs of observers ensured the field observations in the present study. We limited birds' disturbance with two decisions: tests were of short duration, and birds were tested only once. Moreover, after our tests, we checked that all birds returned to their pre-test behavior in less than 5 min.

Statistical Analysis

We followed the same methodology as Dutour et al. (2020) to analyze our results. We therefore split our tests into two questions: first, we compared the response of great tits to natural conspecific and natural allopatric calls (species comparison). Then, we compared responses to the control (background noise), the natural allopatric call, and the reversed allopatric call (order comparison). We used GLMM (*glmer*, package *lme4*) for both the scanning and approach behavior, with the original soundtrack as a random effect. Posthoc comparisons were achieved with functions *emmeans* and *multcomp:cld* (packages *emmeans* and *multcomp*) with a Tukey adjustment for multiple comparisons. The number of scans produced was analyzed with a Poisson distribution and log link function, since no overdispersion was detected (checked with *glmm.overdisp*, package *RVAideMemoire*). Note that Dutour et al. (2020) used a quasi-Poisson distribution because of overdispersion of their data; and that the analysis of Experiment 2

took into account the identity of the bird tested, as they were tested multiple times. We also corrected the analyses for the actual observation time using the time the bird was seen as an offset. For the approach behavior, we set a logistic regression (binomial distribution and logit link function). All fixed effects introduced in the models were tested using Wald tests (*Anova*, package *car*).

Since the raw data available in the supplementary material of Dutour et al. (2020) is incomplete regarding the cross over design used for the second experiment, it was not possible to embed both our dataset and the one of Dutour et al. (2020) in the same analysis in order to compare both studies. Nevertheless, the available information published in Dutour et al. (2020) was sufficient to calculate the effect sizes of each relevant comparison, and we therefore used these metrics to compare our results to those of Dutour et al. (2020). We computed odds ratio (hereafter OR, *odds.ratio*, package *questionr*) for the approach behavior, and Cliff's d for the scanning behavior as this variable does not follow a normal distribution (*cliff.delta*, package *effsize*). One should nevertheless note that the computed effect size does not take into account the non-independence of the observations done in the second experiment of Dutour et al. (2020; i.e., the cross-over design where different acoustic tests were performed on the same subjects).

Results

Question 1: Response to Natural Mobbing Calls from a Conspecific or an Allopatric Species

In our experiment, great tits scanned an average of 7.30 ± 3.16 scans (mean ± standard deviation) when presented with conspecific calls, and 6.80 ± 3.12 scans when presented with black-capped chickadee calls (Figure 2). No statistical difference was detected in our model ($\chi^2 = 0.93$, df = 1, $p = .33$), and the calculated effect size of the difference was 0.18 (Cliff's d, 95% CI [-0.19; 0.51], Table 2). In Dutour et al. (2020), great tits produced 10 ± 5.33 scans in response to conspecific calls and 9.05 ± 5.62 scans in response to black-capped chickadee calls. The resulting effect size is 0.12 (Cliff's d, 95% CI [-0.22; 0.44], Table 2), hence very similar to the one we detected (Figure 3).

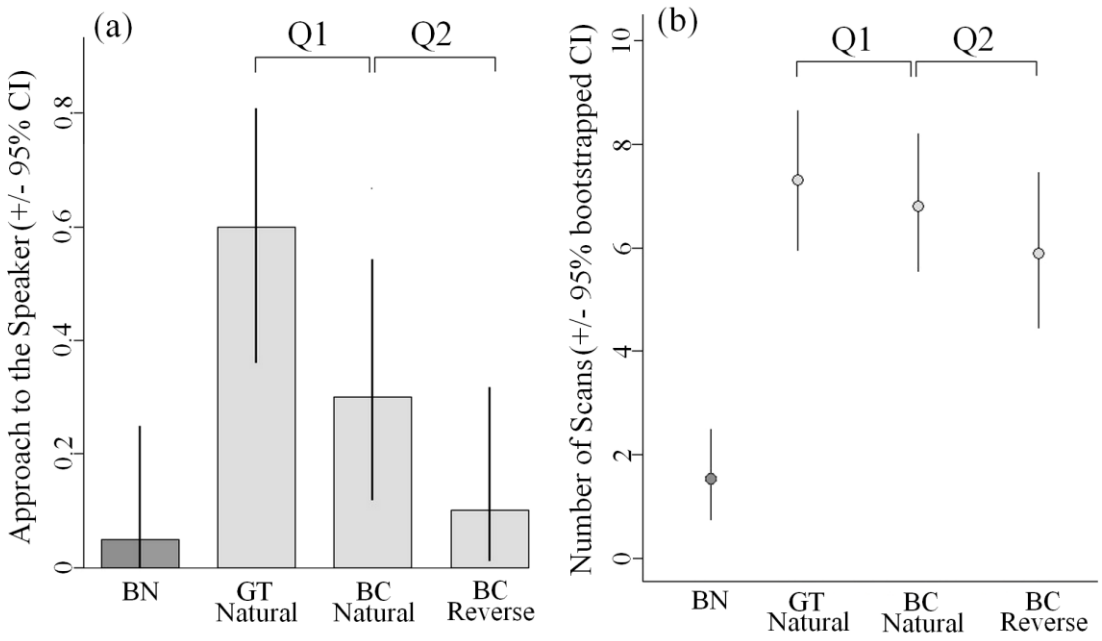
In our experiment, 60% of the great tits tested (n = 20 for each treatment) approached the loudspeaker when hearing conspecific calls, but only 30% when hearing black-capped chickadee calls (Figure 2), and the difference between both treatments approached statistical significance ($\chi^2 = 3.51$, df = 1, $p = .06$), with an odds ratio of 3.5 (95% CI [0.94; 12.97]). This difference was stronger than in Dutour et al. (2020), who found an odds ratio of 1.5 (95% CI [0.42; 5.24], Table 2) between the two treatments. Nonetheless, the confidence intervals of the effect sizes being large (Cumming et al., 2007), the difference between our two studies cannot be considered as statistically significant (Figure 3).

Question 2: Response to Reversed Allopatric Calls

In our experiment, great tits scanned differently background noise, natural allopatric calls, and reversed allopatric calls ($\chi^2 = 56.04$, df = 2, $p < .001$, Figure 2). Indeed, they scanned less to the background noise than to either of the two allopatric soundtracks (BN vs BC Natural: $z = 7.41$, $p < .001$; BN vs BC Reversed: $z = 6.59$, $p < .001$, Figure 2). They produced, on average, 6.8 ± 3.12 scans toward the natural calls, and 5.9 ± 3.54 scans toward the reversed, leading to an effect size of 0.16 (95% CI [-0.20; 0.48], Table 2), which is a non-statistically significant difference as indicated by post-hoc tests ($z = 1.15$, $p = .48$). Birds only produced 1.55 ± 2.06 scans when hearing control tests. In contrast, in Dutour et al. (2020), great tits scanned on average 14.3 ± 6.80 scans to the natural calls, 11 ± 6.55 scans to the reversed calls, and 8.85 ± 6.33 the control tests, leading to a substantial difference between natural and reversed calls (0.27, 95% CI [-0.09; 0.58]) and no significant difference between reversed calls and background noise (0.21, 95% CI [-0.15; 0.52]). Nonetheless, the effect sizes associated to these differences remain comparable between the two studies (Figure 3).

Figure 2

Proportion of Individuals that Approached the Loudspeaker When Hearing the Different Treatments in our Experiment (a), and Number of Scans they Produced (b)



Note. For both figures, the 95% confidence intervals are given. Q1 represents the comparison of interest for the first question (emitter species comparison), comparable to the experiment 1 of Dutour et al. (2020). Q2 represents the comparison of interest for the second question (order comparison), comparable to the experiment 2 of Dutour et al. (2020). Statistical inference can be made using the overlap of such CI: if they overlap at less than halfway, the difference can be considered as statistically significant for an alpha = 5% (Cumming et al., 2007). BN = Background noise, GT = great tit, BC = black-capped chickadee.

Table 2

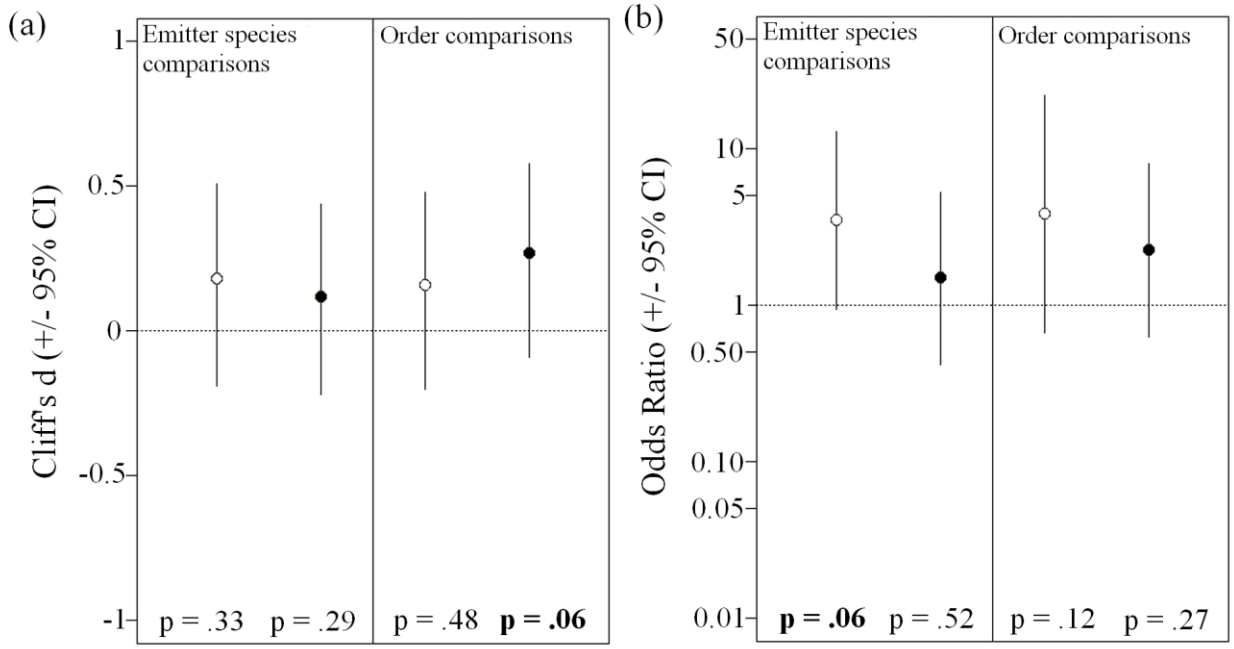
Comparisons of P-Value and Effect Sizes of Both Studies (Salis et al. or Dutour et al.) Regarding the Differences Between Treatments

		Salis et al. (current paper)		Dutour et al. (2020)	
Comparison	Behavior	Conclusion based on p-value	Effect size	Conclusion based on p-value	Effect size
BC-Natural vs GT-Natural	Approach	Marginal effect (p = .06)	3.5 [0.94; 12.97]	No difference (p = .52)	1.5 [0.42; 5.24]
	Scan	No Difference (p = .33)	0.18 [-0.19; 0.51]	No difference (p = .29)	0.12 [-0.22; 0.44]
BC-Natural vs BC-Reversed	Approach	No Difference (p = .12)	3.85 [0.67; 22.11]	No difference (p = .27)	2.26 [0.63; 8.10]
	Scan	No Difference (p = .48)	0.16 [-0.20; 0.48]	Marginal effect (p = .06)	0.27 [-0.09; 0.58]
BC-Reversed vs Control	Approach	No difference (p = .80)	0.47 [0.04; 5.69]	No difference (p = .32)	0.33 [0.07; 1.52]
	Scan	Difference (p < .001)	0.73 [0.43; 0.89]	No difference (p = .35)	0.21 [-0.15; 0.52]

Note. Effect sizes from the scanning behavior are Cliff's d given with their 95% confidence intervals (CI). If such CI encompass 0, the difference can be considered as non-statistically significant. Effect sizes from the approach behavior are Odds Ratio given with their 95% CI. If such CI encompass 1, the difference can be considered as non-statistically significant.

Figure 3

Comparison of Effect Sizes Between Our Own Experiment (White Dots) and Results from Dutour et al. (2020, Black Dots) Who Tested the Same Population With the Same Treatments



Note. (a) Represents the comparisons for the scanning variable, using Cliff's D, and (b) the comparisons for the approach variable, with odds ratio. For each effect size given, the associated 95% confidence intervals are given. For both (a) and (b), the left part concerns question 1 (species comparison: GT-Natural vs BC-Natural) and the right part concerns question 2 (order comparison: BC-Natural vs BC-Reversed). Statistical inference can be made using the overlap of such CI: if they overlap at less than halfway, the difference can be considered as statistically significant for an alpha = 5% (Cumming et al., 2007). Associated *p*-values found in respective models are indicated below each comparison.

In our experiment, only 5% of great tits approached the loudspeaker when hearing background noise, but 30% approached when hearing naturally ordered allopatric calls and 10% approached when hearing reversed allopatric calls (Figure 2). Even though the odds ratio of the difference between our treatments was superior to 1 (Figure 3), it was not statistically significant ($\chi^2 = 4.29$, df = 2, $p = .12$). Our effect sizes parallel the ones from Dutour et al. (2020) who also did not detect statistically significant difference between natural and reversed calls (Table 2, Figure 3). Nonetheless, the percentage of approached in Dutour et al. (2020) was overall higher, with 55% of birds approaching in response to natural allopatric calls, 35% for the reversed calls, and 15% for BN.

Discussion

Two researchers with the same idea, very similar protocols and statistical analyses obtained similar effect sizes for the differences of interest, but nonetheless differed in their conclusions on the great tits' communicative abilities when they were based on the *p*-value. Indeed, we detected a lower response to black-capped chickadee calls compared to conspecific ones, while the responses to both calls were similar in Dutour et al. (2020). We detected no difference between responses to natural and reversed allopatric calls while Dutour et al. (2020) detected one for scanning. Whereas the difference between great tits' and black-capped chickadees' natural calls can easily be explained by a subtle protocol choice; the difference regarding the second question (i.e., effect of reversion on great tits' response) could be explained both by a protocol choice and/or by the *p*-value fluctuating especially with low sample sizes (N

= 20 for each treatment in both studies). These two disparities are therefore of different kinds and will be discussed below.

Allopatric Versus Conspecific Mobbing Calls

In our experiment, great tits approached less to allopatric calls than to conspecific ones, a result different from Dutour et al. (2020) who did not detect any difference. A lower response from great tits to black-capped chickadee mobbing calls has previously been detected in Randler (2012), while a similar level of response was found in Dutour et al. (2017). One could hypothesize that such difference is explained by the distance of the loudspeaker from the focal bird (30 m for Dutour et al., 2020 versus 16 m for us). Indeed, amplitude of the sound is probably a proxy for urgency in birds (Hingee & Magrath, 2009) and calls uttered at larger distance could consequently engender lower approach. In addition, increased distance implies both the attenuation of the sound (lower sound to noise ratio) and the degradation of some sound characteristics (e.g., high frequencies are degraded more easily, Kroodsma et al., 1982). Sound attenuation and degradation have been repeatedly shown to modify birds' response, especially in studies investigating anthropogenic noise (Jung et al., 2020; Shannon et al., 2016). In our situation, the differences between the mobbing calls of the allopatric black-capped chickadee and the sympatric marsh tit (that possess a similar mobbing call) could therefore be less salient at longer distances. However, two points should be raised: firstly, sound attenuation and degradation of a sound at 30 m (Dutour et al., 2020) versus 15 m (our experiment) in a semi-open environment is probably extremely similar. Secondly, the discriminative skills of parids are known to be particularly precise. For example, black-capped chickadees and mountain chickadees (*Poecile gambeli*) can distinguish each other's calls based on features of their D notes (Bloomfield et al., 2008).

We rather suggest that such differences may lie in the soundtrack preparation, and particularly in the number of D notes. Indeed, D notes possess a general recruitment function in some species of parids (Dutour, Lengagne et al., 2019; Suzuki et al., 2016) and the number of D notes per call is thought to code for urgency in parids (Kalb et al., 2019b; Templeton et al., 2005). In our case, because black-capped chickadees' notes are longer than great tits' notes, each researcher chose to either control for call length or for the number of D notes per call. Dutour et al. (2020) chose to control the number of D notes, with 8 notes per call in all playbacks, while we chose to control the call length resulting in only 2 or 3 D notes per call for the black-capped chickadees' playbacks versus 6-8 D notes for great tits' playbacks (Figure 1). The duty cycles (i.e., the amount of time a signal is present over a specified time, Landsborough et al., 2019) between our treatments were overall similar (20 to 30s/ min of playbacks), because Dutour and colleagues chose to reduce the number of call repetition for the black-capped chickadees' playbacks. Future experiments disentangling the effect of D note number versus number of call repetitions with a crossed design may be of interest. Importantly however, even if the response to BC calls was lower in our experiment, the effect sizes of the differences between natural and reversed order in our second question were similar in Dutour et al. (2020) and our own experiment: different choices in protocol did not hamper subsequent differences of interest.

Difference in Scan Number

Our second disparity lies in the difference in scanning behavior for the second question. The absolute number of scans was extremely different, with rarely more than 10 scans counted in our study, while most observations from Dutour and colleagues counted more than 10 scans. In addition, the difference between control and reversed BC playbacks was strong in our study, but not significant in Dutour et al. (2020). The scanning variable could be criticized: counting 180° head turn in real time may be difficult and is probably impacted by the observer's personal definition of scanning. However, the two observers in our study varied only in their scan number for 1 scan on average. Such a result is in accordance with Dutour, Lengagne et al. (2019b) who tested the differences in scan count between two experienced ornithologists and detected a high concordance between observers. The difference in absolute

scores between our two studies may consequently rather be explained by the context in which the birds were tested. Indeed, while we tested free ranging birds while foraging, Dutour and colleagues tested birds when arriving at their nest box. Birds are probably more vigilant (hence increasing the number of scan) in the vicinity of their nest, and the perceived risk associated to conspecific and allopatric calls could also differ according to the distance of the caller from the nest. This subtle variation of context between both studies could thus well explain both the stronger difference between BN and Reversed playback in our study compared to Dutour et al. (2020), and the overall disparities of the absolute scores between the two studies. A question that remains is whether, in addition to difference in absolute scores of scanning, such difference in context may also affect the differences between treatments. Since the effect sizes of the differences between treatments were similar between our experiment and the one from Dutour et al. (2020), we think that the context overall increased the scanning behavior but did not affect the differences between treatments.

Similar Effect Sizes, but Dissimilar *p*-values

Obtaining similar effect sizes of the difference between natural and reversed calls indicates two important things. Firstly, this indicates that even if Dutour et al. (2020) were not fully blinded when doing their playback tests, they were not affected by an expectancy effect (i.e., unknowingly distorting the observations to make them fit with the hypothesis, Holman et al., 2015; Rosenthal & Fode, 1963). Secondly, obtaining similar effect sizes but dissimilar *p*-values between the two studies indicates a discrepancy between effect sizes and analyses based on *p*-values. The use of *p*-value is increasingly criticized (Anderson et al., 2000). Indeed, *p*-values are known to fluctuate even with great sample sizes (Halsey et al., 2015). In our case, natural variability combined with the difference of experimental design between both studies (i.e., completely random versus partly cross over design) could well have contributed to this phenomenon. Indeed, the slightly lower *p*-value reported by Dutour et al. (2020) may have arisen from a higher statistical power of the cross over design permitted by the subtraction of the predicted individual variability from the residual variance (i.e., through the inclusion of a random individual effect). Unfortunately, the estimate of the subject effect was not reported in Dutour et al. (2020) precluding the possibility to examine this point more formally. Our results emphasize the need to report effect sizes and to refer to their biological relevance (Nakagawa & Cuthill, 2007), especially since they seem more stable than *p*-values with low sample sizes (Halsey et al., 2015). Clustering several 'mini-experiments' may be another solution to control natural and protocol variability (von Kortzfleisch et al., 2020). More generally, the various sources of variability between the two experiments (protocol choices and natural between-year variability) show how much replicated studies and meta-analysis approaches are needed.

Conclusion

An accidental situation generated a unique opportunity to compare two studies without adding a null-hacking bias into the replication process. We found that the context in which the birds are tested (here, different distances from the nest) as much as the playback preparation can modify the behavioral cues assessed in language related studies. These different protocol choices seem to have mainly affected the absolute scores rather than the differences between treatments, as we found similar effect sizes between the two experiments. However, relying only on the *p*-value would here have led to different biological conclusions regarding complex syntax use in great tits. In our field of research, the flexibility present in protocol choices and the limited sample sizes are probably the major explanations for disparities between similar experiments. We believe this work provides a clear demonstration that discrepancies in findings between similar experiments should not be taken as the sign that certain methodologies are inherently flawed. Rather, replications should be regarded as a great opportunity to estimate variability in natural experiments and understand how robust a finding can be, with the final aim of approaching, at best, biological reality.

Acknowledgements

We thank Mylène Dutour, Toshitaka Suzuki and David Wheatcroft for their transparency, allowing us this comparison study. Collaborations with M. Dutour on other projects are currently being reviewed but the current study was done independently from her work. We thank the Fondation Vérots for access on their property. Finally, we thank Charlotte Bourbon, Jean Capelle and Julie Ruffion for useful help in the field.

The authors complied with the ASAB guidelines for the use of animals in research. The fieldwork did not require any special permit but followed the laws of the Rhône county and the rules of the ethics committee of the University Lyon 1.

References

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *The Journal of Wildlife Management*, *64*, 912-923.

Baker, M. C., & Becker, A. M. (2002). Mobbing calls of black-capped chickadees: Effects of urgency on call production. *The Wilson Bulletin*, *114*, 510-516.

Bloomfield, L. L., Farrell, T. M., & Sturdy, C. B. (2008). All "chick-a-dee" calls are not created equally Part II. Mechanisms for discrimination by sympatric and allopatric chickadees. *Behavioural Processes*, *77*, 87-99.

Bohannon, J. (2015). Many psychology papers fail replication test. *Science*, *349*, 910-911.

Bolhuis, J. J., Beckers, G. J. L., Huybregts, M. A. C., Berwick, R. C., & Everaert, M. B. H. (2018a). Meaningful syntactic structure in songbird vocalizations? *PLOS Biology*, *16*, e2005157.

Bolhuis, J. J., Beckers, G. J. L., Huybregts, M. A. C., Berwick, R. C., & Everaert, M. B. H. (2018b). The slings and arrows of comparative linguistics. *PLOS Biology*, *16*, e3000019.

Bryan, C. J., Yeager, D. S., & O'Brien, J. M. (2019). Replicator degrees of freedom allow publication of misleading failures to replicate. *Proceedings of the National Academy of Sciences*, *116*, 25535-25545.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365-376.

Carlson, N. V., Healy, S. D., & Templeton, C. N. (2018). Mobbing. *Current Biology*, *28*, R1081-R1082.

Carlson, N. V., Pargeter, H. M., & Templeton, C. N. (2017). Sparrowhawk movement, calling, and presence of dead conspecifics differentially impact blue tit (*Cyanistes caeruleus*) vocal and behavioral mobbing responses. *Behavioral Ecology and Sociobiology*, *71*, 1-11.

Cumming, G., Fidler, F., & Vaux, D. L. (2007). Error bars in experimental biology. *Journal of Cell Biology*, *177*, 7-11.

Danchin, E., Giraldeau, L. A., & Cézilly, F. (2008). *Behavioural ecology*. Dunod.

Dutour, M., Léna, J.-P., Dumet, A., Gardette, V., Mondy, N., & Lengagne, T. (2019). The role of associative learning process on the response of fledgling great tits (*Parus major*) to mobbing calls. *Animal Cognition*, *22*, 1095-1103.

Dutour, M., Léna, J.-P., & Lengagne, T. (2017). Mobbing calls: A signal transcending species boundaries. *Animal Behaviour*, *131*, 3-11.

Dutour, M., Lengagne, T., & Léna, J. (2019). Syntax manipulation changes perception of mobbing call sequences across passerine species. *Ethology*, *125*, 635-644.

Dutour, M., Suzuki, T. N., & Wheatcroft, D. (2020). Great tit responses to the calls of an unfamiliar species suggest conserved perception of call ordering. *Behavioral Ecology and Sociobiology*, *74*, 37.

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, *115*, 2628-2631.

Géroudet, P., & Robert, L. P. (Eds.). (1954). *Les passereaux: II. des mésanges aux fauvettes*. Delachaux & Niestlé.

Griesser, M., Wheatcroft, D., & Suzuki, T. N. (2018). From bird calls to human language: Exploring the evolutionary drivers of compositional syntax. *Current Opinion in Behavioral Sciences*, *21*, 6-12.

Grimes, D. R., Bauch, C. T., & Ioannidis, J. P. A. (2018). Modelling science trustworthiness under publish or perish pressure. *Royal Society Open Science*, *5*, 171511.

Hailman, J. P., Ficken, M. S., & Ficken, R. W. (1985). The 'chick-a-dee' calls of *Parus atricapillus*: A recombinant system of animal communication compared with written English. *Semiotica*, *56*, 191-224.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle *P* value generates irreproducible results. *Nature Methods*, *12*, 179-185.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of *p*-hacking in science. *PLOS Biology*, *13*, e1002106.

Higginson, A. D., & Munafò, M. R. (2016). Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLOS Biology*, *14*, e2000995.

Hingee, M., & Magrath, R. D. (2009). Flights of fear: A mechanical wing whistle sounds the alarm in a flocking bird. *Proceedings of the Royal Society B: Biological Sciences*, *276*, 4173-4179.

Holman, L., Head, M. L., Lanfear, R., & Jennions, M. D. (2015). Evidence of experimental bias in the life sciences: Why we need blind data recording. *PLOS Biology*, *13*, e1002190.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124.

Jung, H., Sherrod, A., LeBreux, S., Price, J. M., & Freeberg, T. M. (2020). Traffic noise and responses to a simulated approaching avian predator in mixed-species flocks of chickadees, titmice, and nuthatches. *Ethology*, *126*, 620-629.

Kalb, N., Anger, F., & Randler, C. (2019a). Great tits encode contextual information in their food and mobbing calls. *Royal Society Open Science*, *6*, 191210.

Kalb, N., Anger, F., & Randler, C. (2019b). Subtle variations in mobbing calls are predator-specific in great tits (*Parus major*). *Scientific Reports*, *9*, 6572.

Krebs, J. R. (1971). Territory and breeding density in the great tit, *Parus major* L. *Ecology*, *52*, 2-22.

Kroodsma, D. E., Miller, E. H., & Ouellet, H. (Éds.). (1982). *Acoustic communication in birds*. Academic Press.

Landsborough, B., Wilson, D. R., & Mennill, D. J. (2019). Variation in chick-a-dee call sequences, not in the fine structure of chick-a-dee calls, influences mobbing behaviour in mixed-species flocks. *Behavioral Ecology*, *31*, 54-62.

Lash, T. L., Collin, L. J., & Van Dyke, M. E. (2018). The replication crisis in epidemiology: Snowball, snow job, or winter solstice? *Current Epidemiology Reports*, *5*, 175-183.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*, 487-498.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews*, *82*, 591-605.

Otter, K. A. (Ed.). (2007). *The ecology and behavior of chickadees and titmice: An integrated approach*. Oxford University Press.

Randler, C. (2012). A possible phylogenetically conserved urgency response of great tits (*Parus major*) towards allopatric mobbing calls. *Behavioral Ecology and Sociobiology*, *66*, 675-681.

Rosenthal, R., & Fode, K. L. (1963). The effect of experimenter bias on the performance of the albino rat. *Behavioral Science*, *8*, 183-189.

Salis, A., Léna, J., & Lengagne, T. (2021). Great tits (*Parus major*) adequately respond to both allopatric combinatorial mobbing calls and their isolated parts. *Ethology*, *127*, 213-222.

Schmidt, F. L., & Oh, I.-S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology*, *4*, 32-37.

Schwagmeyer, P. L., & Mock, D. W. (1997). How to minimize sample sizes while preserving statistical power. *Animal Behaviour*, *54*, 470-474.

Shannon, G., McKenna, M. F., Angeloni, L. M., Crooks, K. R., Fristrup, K. M., Brown, E., Warner, K. A., Nelson, M. D., White, C., Briggs, J., McFarland, S., & Wittemyer, G. (2016). A synthesis of two decades of research documenting the effects of noise on wildlife: Effects of anthropogenic noise on wildlife. *Biological Reviews*, *91*, 982-1005.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I.,… Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*, 337-356.

Suzuki, T. N., Griesser, M., & Wheatcroft, D. (2019). Syntactic rules in avian vocal sequences as a window into the evolution of compositionality. *Animal Behaviour*, *151*, 267-274.

Suzuki, T. N., Wheatcroft, D., & Griesser, M. (2016). Experimental evidence for compositional syntax in bird calls. *Nature Communications*, *7*, 10986.

Suzuki, T. N., Wheatcroft, D., & Griesser, M. (2020). The syntax–semantics interface in animal vocal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*, 20180405.

Tang-Martínez, Z. (2020). The history and impact of women in animal behaviour and the ABS: A North American perspective. *Animal Behaviour*, *164*, 251-260.

Templeton, C. N., Greene, E., & Davis, K. (2005). Allometry of alarm calls: Black-capped chickadees encode information about predator size. *Science*, *308*, 1934-1937.

von Kortzfleisch, V. T., Karp, N. A., Palme, R., Kaiser, S., Sachser, N., & Richter, S. H. (2020). Improving reproducibility in animal research by splitting the study population into several 'mini-experiments.' *Scientific Reports*, *10*, 16579.

Wilkin, T. A., Garant, D., Gosler, A. G., & Sheldon, B. C. (2006). Density effects on life-history traits in a wild population of the great tit *Parus major*: Analyses of long-term data with GIS techniques: Great tit breeding density. *Journal of Animal Ecology*, *75*, 604-615.